DOCUMENT RESUME

ED 266 288 CE 043 577

AUTHOR Swartz, Richard; And Others
TITLE Adding an Essay to the GED Writing Skills Test: Reliability and Validity Issues. GED Testing Service Research Studies, Number 7.
INSTITUTION American Council on Education, Washington, D.C. General Educational Development Testing Service.
REPORT NO ISBN-0-8268-1203-6
PUB DATE Jul 85
NOTE 20p.; For a related document, see CE 043 578.
AVAILABLE FROM GED Testing Service, American Council on Education, One Dupont Circle, NW, Suite 20, Washington, DC 20036 ($3.50).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Statistical Data (110)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Essay Tests; *High School Equivalency Programs; *Scoring; *Test Reliability; *Test Validity; *Writing Evaluation; Writing Research; Writing Skills
IDENTIFIERS *General Educational Development Tests

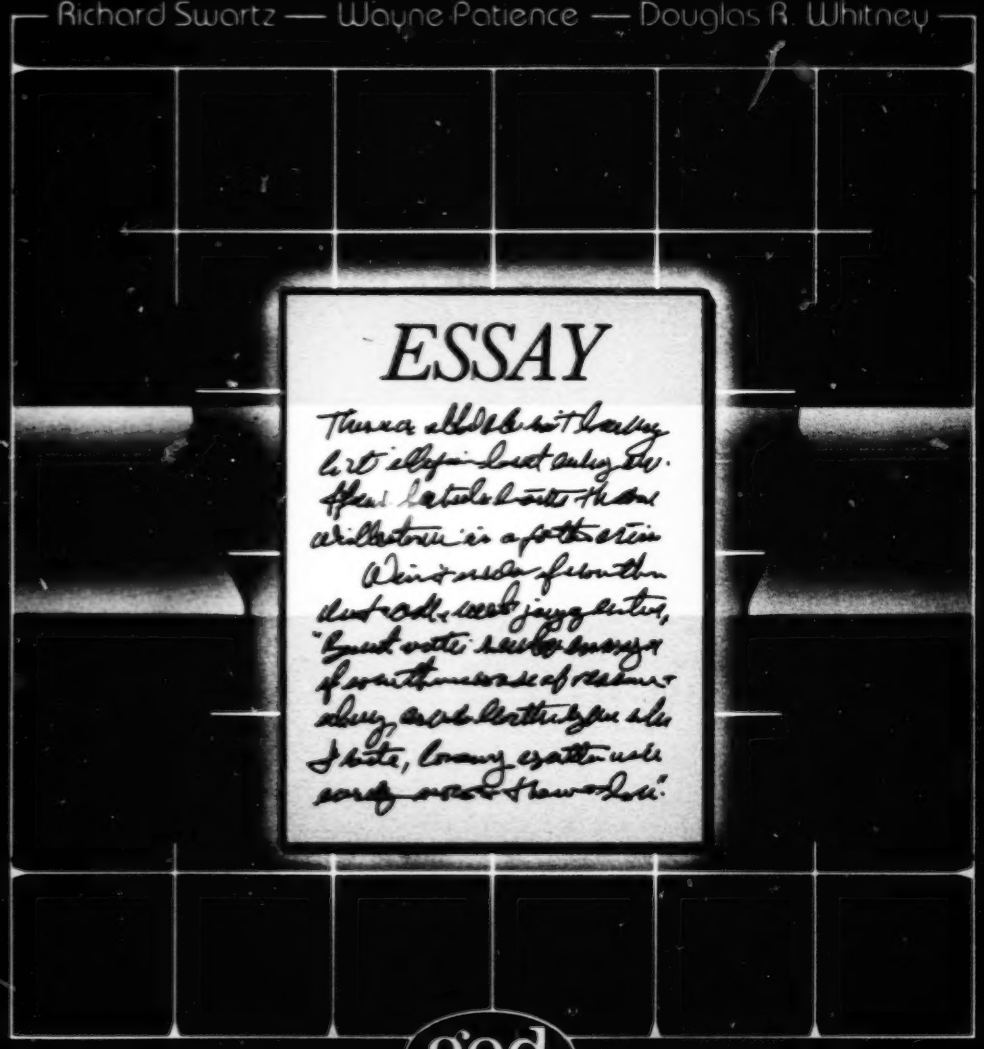
ABSTRACT

In preparation for adding an essay test to the General Educational Development (GED) test, the GED Testing Service undertook a series of studies to establish (1) whether acceptable reading reliabilities were attainable in decentralized holistic scoring sessions often involving no more than a dozen papers; (2) whether essay readers in a variety of locations, employing a variety or procedures, could consistently apply the same essay scoring standards (i.e., achieve scale stability); and (3) whether an essay would measure aspects of writing performance different from those currently measured by the multiple-choice test. The research studies involved the holistic scoring of essays written by more than 2,000 GED examinees from Washington, D.C., Iowa, and Maryland, and by a national sample of almost 4,000 high school seniors. Results indicate that acceptable levels of reading reliability are possible in decentralized scoring sessions. However, two studies of essay scoring scale stability yielded different results, demonstrating the need for additional monitoring techniques during scoring sessions to ensure that the score scale remains constant. To determine whether an essay represents a measure clearly different from the multiple-choice writing test, standard scores on the current GED Writing Skills (multiple-choice) test were correlated with holistic scores on an essay test. The extent of these correlations indicates that the two tests are measuring related but clearly different set of skills. The GED Testing Service plans to conduct additional research to identify the most effective methods for maintaining scoring scale stability in essay scoring sessions conducted in a variety of settings and to ensure that scores earned by students on different essay topics are comparable. (Author/KC)

ED266288

ADDING AN ESSAY TO THE GED WRITING SKILLS TEST: RELIABILITY AND VALIDITY ISSUES

Richard Swartz — Wayne Patience — Douglas R. Whitney



ged

GED TESTING SERVICE
RESEARCH STUDIES, NUMBER 7
JULY 1985

CE 043577

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D Whitney

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The American Council on Education (ACE) founded in 1918 and composed of institutions of higher education and national and regional education associations, is the nation's major coordinating body for postsecondary education. Through voluntary and cooperative action, the Council provides comprehensive leadership for improving education standards, policies, procedures, and services.

The GED testing program began in 1942 with the development of a battery of tests for United States military personnel who had not completed their high school studies. Today, through the GED Testing Service of the American Council on Education, the test are administered in all fifty states, in many Canadian provinces, and in the U.S. commonwealths and territories. The GED testing program has offered millions of adults a "second chance" by enabling them to demonstrate that they have many of the skills they would have acquired had they been able to remain in high school.

Copies of this report are available for $3.50 (prepaid) from the GED Testing Service (Research and Development), American Council on Education, One Dupont Circle, Suite 20, Washington, D.C. 20036.

©1985 by American Council on Education

ISBN: 0-8268-1203-6

ADDING AN ESSAY TO THE GED WRITING SKILLS TEST: RELIABILITY AND VALIDITY ISSUES

Richard Swartz
Wayne Patience
Douglas R. Whitney

GED TESTING SERVICE
RESEARCH STUDIES, NUMBER 7
JULY 1985

AMERICAN COUNCIL ON EDUCATION

Acknowledgements

The research projects described in this report would not have been possible without the enthusiasm and assistance of many individuals. Our special thanks go to John Fields, John Hartwig, Sandy Havran, and Alex Hyman, GED State Administrators from Maryland, Iowa, New Jersey, and Washington, D.C., respectively. The chief examiners in those four locations deserve credit, too, for their willingness to devote the extra time and attention necessary for the administration of an essay test to GED examinees. In Iowa particularly, many other people participated enthusiastically, from adult education teachers who introduced writing into the curriculum to educators who worked hard to make the holistic scoring system work in a local setting. Finally, to the GED examinees in these four locations, and to the high school seniors who participated in the national equating study, we owe a sincere thanks for devoting their time and energy to a task that, alas, many found all too alien.

Table of Contents

ABSTRACT 4

BACKGROUND 5

RELIABILITY 5

CONTENT VALIDITY 10

CONCURRENT VALIDITY 11

CONCLUSIONS 12

REFERENCES 13

APPENDIX A: GED Essay Scoring Guide 14

APPENDIX B: How the GED Essay is Scored 15

Abstract

In February 1984, the General Educational Development (GED) Testing Service appointed a GED Tests Specifications Committee to make recommendations regarding the content of the GED Tests to be published for use in 1988 and beyond. In its report, the Specifications Committee recommended that the GED Writing Skills test, currently a multiple choice item exam, include an essay component.

The GED Testing Service undertook a series of studies to establish 1) whether acceptable reading reliabilities were attainable in decentralized holistic scoring sessions often involving no more than a dozen papers; 2) whether essay readers in a variety of locations, employing a variety of procedures, could consistently apply the same essay scoring standards (i.e., achieve scale stability); and 3) whether an essay would measure aspects of writing performance different from that measured by the current multiple choice test. The research studies involved the holistic scoring of essays written by over 2,000 GED examinees from Washington, D.C., Iowa, and Maryland and by a national sample of almost 4,000 high school seniors.

Results indicated that acceptable levels of reading reliability are possible in decentralized scoring sessions. Readers awarded the same or contiguous scores for 93% of the essays. This rate matched that typical of centralized scoring sessions and yielded a reading reliability of .82. Two studies of essay scoring scale stability yielded different results. In the first study, comparisons of scores given by readers in two different locations indicated that the average scores were not significantly different, demonstrating that the same scale was applied in both locations. In a replication of this analysis with two other scoring sites, however, a significant difference was found, demonstrating the need for additional monitoring techniques during scoring sessions to ensure that the score scale remains constant. To determine whether an essay represented a measure clearly different from the multiple choice writing test, standard scores on the current GED Writing Skills (multiple choice) test were correlated with holistic scores on an essay test. The correlations ranged from .48 for GED examinees to .69 for high school seniors. The extent of these correlations indicates that the two tests are measuring related but clearly different sets of skills. The higher correlation for the high school population probably reflects that population's greater familiarity and facility with essay writing tasks.

The GED Testing Service plans to conduct additional research to identify the most effective methods for maintaining scoring scale stability in essay scoring sessions conducted in a variety of settings. A second research project involves a study of methods for ensuring that scores earned by students on different essay topics are comparable.

Adding An Essay to the GED Writing Skills Test: Reliability and Validity Issues

Richard Swartz • Wayne M. Patience • Douglas R. Whitney

In the course of conducting research on the possible inclusion of an essay in the GED Writing Skills Test, the GED Testing Service (GEDTS) has accumulated a wealth of information regarding direct (essay) and indirect (usually multiple choice) measures of writing. This report summarizes the staff's major findings regarding the technical aspects of adding an essay test to the GED Tests battery. These findings are discussed in terms of their applicability to three critical aspects of test construction: ***reliability, content validity***, and ***concurrent validity***. Additional details regarding the projects described below are available in three other publications: "The Relationship Between Scores on the GED Writing Skills Test and on Direct Measures of Writing" (Swartz & Whitney, 1985); "GED Essay Feasibility Studies: Practical Report" (Swartz, 1985); and "Iowa GED Writing Skills Pilot Project: Final Report" (Iowa Department of Public Instruction, 1985).

Background

The technical basis of the proposal to include an essay as part of the GED Writing Skills test is that an essay will improve the validity of the test. The current GED Writing Skills test, which focuses on usage and editing skills in a multiple choice format, is an indirect measure of writing. Though high degrees of reliability have been demonstrated for multiple choice measures of writing, the validity of these indirect measures has been challenged frequently by writing teachers and administrators (Bamberg, 1982; Charney, 1984; Godshalk, 1966; White, 1984). As Breland (1983) observes, those who favor direct measures of writing have a compelling logic at their disposal: the best way to measure a person's mastery of a skill is to ask for a demonstration of the skill, not for a correlate of it.

As reported in GED Research Brief No. 6, (Whitney, Malizio, & Patience, 1985), the GED Testing Service evaluates the GED Tests for both content validity (Does the test appear to measure the appropriate skills?) and concurrent validity (Do the GED Test scores correlate as expected with scores from other tests which similarly measure educational achievement?). The addition of an essay test to the GED Tests battery is germane to both these aspects of validity. However, an oft-recited truism of the testing profession is that a test must be reliable before it can be valid. Thus, before giving consideration to the increased validity possibly offered by adding an essay test, the staff focused its efforts on establishing the reliability of scoring an essay test.

Reliability

Two aspects of essay tests affect the degree to which a large scale writing assessment is reliable. One is ***reading reliability***—the consistency with which different readers award the same rating to the same essay. The second is ***score reliability***—the extent to which an examinee receives the same score on two similar, equally-demanding essay topics. Data analyses are reported below on these two types of essay reliability.

Several statistical methods may be used in assessing reading reliability. For each set of data presented here, three methods are employed: 1) percent of readers awarding the same scores, contiguous scores (scores differing by one point), and discrepant scores (scores differing by two points or more); 2) correlations of first and second ratings; and 3) comparison of means of first and second ratings. Each of these methods is a way of describing the consistency of two independent ratings within a given reading. "Reading" is used here to refer to the processes in a scoring session (e.g., convening of readers, conducting reader training, obtaining two ratings [on a six point scale] for all essays, resolving any discrepant pair of scores for a given essay by a third rating, and the combination of ratings to yield an essay score for each paper on a 2-12 scale). A fourth method of evaluating reading reliability was employed for a subset of the data presented. This method involved the scoring of some essays in two readings. To evaluate scale stability—the consistency of essay scores obtained on the same essays from the two separate readings—the correlations of scores and comparison of mean scores for the readings are reported. For this fourth method, analyses are conducted on a 2-12 scale since the unit of concern is the total essay score from two readings.

Since the fall of 1983, the GED Testing Service has administered experimental essay tests to two groups of high school seniors and three groups of GED examinees, as outlined below.

Population	Dates Tested	Number of Examinees
High school seniors	10/83 to 11/83	2335
High school seniors	2/85 to 3/85	3787
GED Examinees		
Washington, D.C	9/84 to 5/85	663
Iowa	2/85 to 5/85	501
Maryland	4/85 to 6/85	927

The first group of high school senior essays was collected during the fall 1983 equating study. The population from which the study group was drawn consisted of all United States schools offering a 12th grade, exclusive of "special schools" (e.g., schools for the handicapped, penal institutions, etc.), stratified by public and private, estimated grade 12 enrollment, geographic region, and socioeconomic status. The sample included 170 schools. A full description of this project and its results may be found in Swartz & Whitney (1985). In the study, high school seniors wrote essays on three different topics, producing approximately 800 essays on each topic. These papers were scored holistically (Conlan, 1976; Lloyd-Jones, 1977) by three different groups of readers (one group for each of the three topics) using a six point scale. Each essay was scored independently by two readers, and essays receiving scores two or more points apart were scored by a third reader. The total essay score was the combined score of the two readers; for essays scored by three readers, the total score was twice the average of the three scores (rounded to the nearest whole number). Percentage of reader agreement and correlations between the first and second ratings are reported in Table 1. These reliability figures are considered well within acceptable standards (Breland, 1983; Coffman, 1971; White, 1984), thus reconfirming that essay tests can be scored by experienced readers in a central location according to a holistic scoring system and yield acceptably reliable results.

GEDTS collected essays from high school seniors again in the spring 1985 equating study according to a sampling plan similar to that used in the 1983 equating study. The 1985 study included a sample of 145 schools. The primary goal of the 1985 project was to field test a variety of essay topics, but the project results bear on the current discussion of essay reliability.

The scoring of the essays from the 1983 study was completed by highly experienced essay readers who scored papers on only one topic. In contrast, the readers who scored the high school essays in the 1985 reading were largely new to the process of holistic scoring; many were adult educators from the Washington, D.C. and Maryland area with some experience or interest in teaching writing but little direct experience in large scale writing assessment. Also, because the primary objective of the project was to evaluate the effectiveness of a variety of essay topics, readers scored papers on as many as five topics, though only the papers on a single topic were scored at any one time.

Table 2 provides the percentage of reader agreement and the reading reliabilities for this scoring session by topic. The large percentage of discrepant scores in this session is attributable to the purpose of the session itself: readers were primarily responsible for selecting the most usable topics from among those they read. The normal training and reinforcement procedures for maintaining reader agreement were not used in this case because scores were not being reported to students and because rates of reader agreement were not among the subjects of utmost importance in the project.

The two studies summarized above served merely to confirm what is widely recognized in the field (Breland, 1983; Cooper, 1977; White, 1984): that trained readers holistically scoring essays in a controlled setting can, indeed, achieve acceptable degrees of reliability. A third study was undertaken in the fall of 1984 to amplify these findings. Important differences in this study included the fact that the population was to consist of GED examinees; that many potential readers had little or no prior experience holistically scoring essays; that some scoring sessions could not be conducted in a "controlled" setting; and that the same essays were to be scored in more than one location to verify that the scoring guides could be used to impose an adequate standardization in scale despite differences in procedures and location.

Table 1
1983 High School Study
Reading Reliability

Topic	N	Mean	Stand. Dev.	Rater Agreement: % Same Score	% Contig. Score	% Discrep. Scores	Correlation Bet. Readers	Reading Reliability
A. Description	813	6.1	2.3	51%	43%	6%	.76	.86
B. Persuasion	747	6.5	2.4	47%	46%	7%	.76	.86
C. Exposition	775	6.6	2.1	47%	46%	7%	.69	.82

Notes:

1. Rater agreement is based on the six-point holistic scale used by each reader.
2. Since the essay score is on the twelve point scale, correlations reported as "Correlation Between Readers" have been stepped-up using the Spearman-Brown formula in order to represent the expected "Reading Reliability."
3. Differences between the mean ratings one and two for topic A, B, and C were 0.03, 0.09, and 0.08 respectively. Because correlations are insensitive to shifts in scale, these differences in means are reported to confirm that no significant shifts in rating scale were present between ratings one and two.

Table 2
1985 High School Study
Reading Reliability

Topic	N	Mean	Stand. Dev.	Rater Agreement: % Same Score	% Contig. Score	% Discrep. Scores	Correlation Bet. Readers	Reading Reliability
A	576	5.8	2.4	36%	48%	16%	.65	.78
B	605	6.1	2.3	37%	47%	16%	.61	.76
*C	496	6.0	2.3	36%	47%	17%	.60	.75
D	186	6.1	2.4	32%	52%	16%	.62	.77
*E	171	6.1	2.3	34%	49%	17%	.69	.82
F	208	5.8	2.5	41%	45%	14%	.71	.83
G	156	6.1	2.3	39%	47%	14%	.67	.80
H	204	6.0	2.3	33%	43%	24%	.55	.70
I	170	6.0	2.4	33%	46%	21%	.60	.75
J	199	6.0	2.4	37%	48%	15%	.70	.82
*K	192	6.1	2.5	33%	50%	17%	.69	.82
*L	158	6.3	2.1	37%	43%	20%	.55	.71
*M	71	6.1	2.4	30%	48%	22%	.53	.69
N	218	5.6	2.4	35%	44%	21%	.56	.72
*O	177	5.8	2.3	27%	52%	21%	.55	.71
All	3787	6.0	2.3	35%	48%	17%	.61	.76

Notes:

1. Rater agreement is based on the six-point holistic scale used by each reader.
2. Since the essay score is on the twelve point scale, correlations reported as "Correlation Between Readers" have been stepped-up using the Spearman-Brown formula in order to represent the expected "Reading Reliability."
3. Topics marked with an asterisk were judged unacceptable for operational use by essays readers.

The first part of this study began with the administration of an essay test to GED examinees in the District of Columbia. Examinees were given a single topic calling for them to take a position on an issue and to support their position with reasons and examples. A time limit of thirty minutes was established for the writing of the essay. The study examined the scores given to 663 essays written by GED examinees in Washington, D.C. These essays were scored periodically at the GEDTS offices by local educators who, for the most part, had little or no prior experience in holistic scoring.

The Washington scoring sessions were "controlled" sessions. That is, large numbers of papers were scored in one sitting, and each scoring session began with an introduction to holistic scoring, a review of the scoring guide (see Appendix A), a scoring of sample papers, and a discussion of standards. When readers began scoring actual examinee papers, their scoring was monitored by a "table leader" to ensure that individual standards did not replace the standards previously defined by the sample papers and by the scoring guide. The reading reliability achieved in these readings was .70, a finding which served to establish that essays written by GED examinees could be scored holistically with acceptable reliability by readers previously inexperienced in holistic scoring.

The second part of this study involved the administration of an essay to GED examinees in Iowa from February, 1985 to May, 1985. In Iowa, unlike the District of Columbia, essays were scored at local centers throughout the state, so there was no opportunity for the kind of centralized scoring sessions that had been used in Washington. Apart from the numerous practical issues that were immediately apparent, there were two technical issues of concern to GEDTS. First, could essays be scored holistically with sufficient reliability in local settings without the built-in, rigorous controls characteristic of a central scoring session? Second, could the same scoring standards be maintained regardless of setting or particular scoring procedure?

Iowa readers were trained in holistic scoring at a workshop conducted by GEDTS staff and were required to qualify by correctly scoring at least 80% of the papers in a qualifying packet. (For a full description of the Iowa project, see "Iowa GED Writing Skills Pilot Project: Final Report," Iowa Department of Public Instruction, 1985). Scoring standards were maintained in both sites through the use of sample papers and a scoring guide prepared and distributed by GEDTS. Essay readers in Iowa, working in pairs in local centers and usually scoring very few papers at any one time, achieved a reading reliability of .82 on a sample of 501 essays.

Table 3 provides means, standard deviations, and reading reliability coefficients for the Washington and Iowa scoring sessions. As indicated in the table, the set of ratings from Washington has less variance (a smaller standard deviation) than the ratings from Iowa, primarily because of a more restricted distribution of scores among the Washington papers (the lower mean and smaller standard deviation for the Washington papers is also evident in the Writing Skills test [multiple choice] standard scores for Washington examinees as well—see Table 4).

This limited variance for the Washington ratings accounts, at least in part, for the difference in reading reliability coefficients between Washington, D.C. and Iowa; as noted above, the percentages of readers giving the same or contiguous scores to papers was virtually identical in both locations. Had the variance of ratings in the Washington, D.C. readings been similar to that observed in Iowa, the reading reliability would have been .82, the same as that observed in the Iowa readings. The same kind of difference in reading reliability coefficients can be observed between the 1983 high school essay reading and the scoring of GED examinee papers (compare Tables 1 and 2 to 3 and 4). Again, the difference is attributable to a larger variance among scores on the high school student papers. This in-

Table 3
GED Examinees: Washington, D.C. and Iowa
Reading Reliability

				Rater Agreement				
Topic	N	Mean	Stand. Dev.	% Same Score	% Contig. Score	% Discrep. Scores	Correlation Bet. Readers	Reading Reliability
D.C.	663	4.3	1.6	44%	49%	7%	.54	.70
Iowa	501	5.8	2.1	53%	40%	7%	.70	.82
Total	1164	5.0	2.0	48%	45%	7%	.67	.80

Notes:

1. Rater agreement is based on the six-point holistic scale used by each reader.
2. Since the essay score is on the twelve point scale, correlations reported as "Correlation Between Readers" have been stepped-up using the Spearman-Brown formula in order to represent the expected "Reading Reliability."
3. Differences between the mean ratings one and two for readings in D.C. and in Iowa were 0.056 and 0.079 respectively. Because correlations are insensitive to shifts in scale, these differences in means are reported to confirm that no significant shifts in rating scale were present between ratings one and two.

Table 4
GED Examinees: Maryland
Reading Reliability

				Rater Agreement				
	N	Mean	Stand. Dev.	% Same Score	% Contig. Score	% Discrep. Scores	Correlation Bet. Readers	Reading Reliability
	927	4.2	2.0	50%	41%	9%	.65	.79

Notes:

1. Rater agreement is based on the six-point holistic scale used by each reader.
2. Since the essay score is on the twelve point scale, correlations reported as "Correlation Between Readers" have been stepped-up using the Spearman-Brown formula in order to represent the expected "Reading Reliability."
3. Differences between the mean ratings one and two for Maryland papers was 0.119. Because correlations are insensitive to shifts in scale, these differences in means are reported to confirm that no significant shifts in rating scale were present between ratings one and two.

terpretation is confirmed by looking at the reading reliability obtained by the same group of Washington, D.C. raters on a sample of 927 essays written by GED examinees in Maryland (compare Washington, D.C. and Maryland data in Tables 3 and 4). The Maryland essays provided a more variable distribution of scores, thus yielding a higher reading reliability even though the rates of reader agreement were slightly lower.

To check the second, more important, aspect of reader reliability—to what extent readers were using the same scoring standards in both locations (scale stability)—351 papers were scored by both readers in Washington, D.C. and in Iowa. Papers from both locations were mixed together and arranged so that readers did not know where the paper had been written, and no previous scores were apparent on the papers. A comparison of ratings given to these papers in both sites indicated that average scores for papers scored in the two different sites were not significantly different from each other ($t = 1.04$, $df = 347$). The correlation between scores on papers rated in both locations was .73.

Because correlations are insensitive to shifts in scale, it is important to note that the mean total essay scores for a common set of papers did not change significantly from one scoring site to another. This is particularly notable in that the papers written in Iowa had higher mean scores than the Washington, D.C. essays. Because Washington, D.C. essay readers were accustomed to scoring papers clustered around the lower end of the score scale, they might have been expected to overcompensate when they encountered papers representing a broader distribution of scores. However, the similarity of mean scores given in Iowa and in Washington, D.C. to the same papers clearly establishes that this did not happen. The slight differences that were apparent between the two groups are not significant in terms of group statistics, but, in some cases, could have practical significance for individual examinees. The differences that were apparent in this comparison suggest the need for additional procedures for monitoring essay scoring at local sites.

Because maintaining the stability of the essay score scale is the central reliability issue in the inclusion of an essay in the GED Writing Skills test, the study of scale stability was replicated. A set of 328 essays written by Maryland examinees were scored in both Washington, D.C. and by readers in New Jersey. The results are presented in Table 6.

Table 5
Means and Standard Deviations for Papers Scored in Two Locations
GED Examinees: Washington, D.C. and Iowa
Topic A

Essay From		Scored In: Iowa Mean	Iowa S.D.	D.C. Mean	D.C. S.D.
Iowa	Essay	5.8	2.0	5.6	1.9
	Multiple Choice	47.9	6.6		
D.C.	Essay	4.2	2.1	4.3	1.7
	Multiple Choice	45.0	7.5		

Note:
Analysis is based on a sample of 351 essays scored by at least 2 readers at each location.

Table 6

GED Examinees: Maryland (Scored in Washington, D.C. and New Jersey)
Reading Reliability

Scoring Site	N	Mean	Stand. Dev.	Rater Agreement: % Same Score	% Contig. Score	%Discrep Scores	Correlation Bet. Raters	Reading Reliability
D.C.	328	4.00	2.01	56.0%	44.0%	0	.81	.90
NJ	328	4.48	2.06	53.0%	45.0%	2%	.77	.87

Notes:

1. Rater agreement is based on the six-point holistic scale used by each reader.
2. Since the essay score is on the twelve point scale, correlations reported as "Correlations Between Readers" have been stepped-up using the Spearman-Brown formula in order to represent the expected "Reading Reliability."
3. Differences between the mean ratings one and two for readings in Maryland and in New Jersey were -0.08 and -0.03 respectively. Because correlations are insensitive to shifts in scale, these differences in means are reported to confirm that no significant shifts in rating scale were present between ratings one and two.
4. Difference in mean scores by scoring location is significant at $p<0.01$ ($t = -6.59$, $d = 327$).

Content Validity

With the fact established that essays can indeed be scored reliably in a variety of settings, it is appropriate to consider how the addition of an essay would affect the validity of the GED Writing Skills test. In the judgment of professionals in the field, and in light of current practices in the states, the addition of an essay component to the GED Writing Skills test would clearly improve the test's content validity. As a part of the five year review of the GED Tests, GEDTS conducted a series of surveys addressing program goals, test content, and test format (Spille, 1984). Among those polled in these surveys were state directors of adult education, GED Administrators, Chief Examiners, and teachers. Among the questions asked was "Should the GED Tests contain a measure of the ability to produce original written communications?" In each group, a clear majority of respondents favored this change.

The views of the GED constituency were reflected in the recommendations of numerous reports regarding the nature of a high school education in America. These reports cited writing— in the form of composing or drafting an original communication— as important both to academic and vocational pursuits. The National Commission on Excellence in Education (1983) recommended that high school students should be able to "write well-organized, effective papers." Among the learning outcomes in English identified by the Project EQuality report (1983) was "the recognition that writing is a process involving a number of elements, including collecting information and formulating ideas, determining their relationships, drafting, arranging paragraphs in an appropriate order and building transitions between them, and revising what has been written." These two recommendations were typical of the suggestions for writing in the high schools made in a host of national reports in the early 1980s (Applebee, 1981; Boyer, 1983; Goodlad, 1983).

The recommendations regarding skills necessary for success on the job and in industry are similar to those from academe. The College Board met with 200 business leaders and educators as a part of Project EQuality (1984) and came away with a clear message: "Lack of competency in writing also limits the effectiveness of communication in the workplace and hurts productivity and employee relations." The Report of the Panel on Secondary School Education for the Changing Workplace (1984) lists various writing competencies in support of the finding that "Employees in many lines of work are called upon to fill out forms, document experiences and procedures, record events, and present their ideas in memorandums, letters, and notes." Likewise, a group of business representatives assembled by the Task Force on Education for Economic Growth (1983) identified as a necessary competency "the ability to organize, select and relate ideas and to outline and develop them in coherent paragraphs."

The above materials were among those included in a comprehensive review of the literature conducted by the GED Tests Specification Committee, convened by the GED Testing Service in February, 1984 (GEDTS, 1984). In its report including recommendations regarding the Writing Skills test, the Language Arts Panel reaffirmed the emphasis of the national reports for the GED population:

> The English specifications panel believes that no one should receive credit for high school equivalency without being asked to demonstrate writing ability directly as well as indirectly. Thus, we recommend that every candidate be asked to write.

A final aspect of content validity is the degree to which actual practices in the nation's high schools correspond to the content of the GED Tests. A recent survey of the use of direct writing measures by Education Commission of the States (1984) clearly illustrates how powerful the trend is now and suggests that it will gain momentum in the late eighties. Based on information provided in the report, it is possible to characterize direct writing assessment in the United States as follows:

- three states currently use the results of an essay test as one high school graduation requirement, and two others will add a similar requirement within the next two years. The volume of GED testing in these five states represented 20% of the total testing volume in the United States and Canada in 1984;
- 22 states currently use direct measures of writing in statewide assessments, and six others will begin within a year
- by October of 1985, 71% of the people taking the GED Tests in the United States will be in states that conduct direct writing assessments of their public school students in grades 9, 10, 11, or 12.

In sum, adult education policy-makers and practitioners favor the inclusion of an essay; virtually all of the recent national reports on high school education—from vocational as well as academic orientation—stress the importance of the ability to write original communications; a committee of national curriculum experts convened by the GED Testing Service considered an essay an essential part of any test called "writing" and current practices in the states indicate widespread and growing use of direct writing assessments, both for curriculum planning and evaluation as well as for a graduation requirement.

Table 7
Correlations Between Essay Scores For Persons Writing Two Essays (1983 High School Senior Population)

Topic	Topic A	Topic B	Topic C
A. Description	---	.65	.54
B. Persuasion	181	---	.60
C. Exposition	186	185	---

Notes:
1. Correlations are shown above the diagonal; sample sizes below the diagonal.
2. All tabled correlations are significantly different from zero when tested at a 5% significance level.

Concurrent Validity

While the testimony from leaders in education and business throughout the country is clear and consistent regarding the importance of actual writing, another critical question remains: will the addition of an essay component in the Writing Skills test measure skills that are different from the skills measured by the current multiple choice Writing Skills test?

To answer this question, GEDTS computed the correlation between scores earned on the current (multiple choice) GED Writing Skills test and scores earned on an essay component, as scored holistically by two readers using a six point scale. It bears repeating here that these correlation coefficients assume some value only if the total essay scores given by readers can be considered sufficiently reliable. As detailed in the section above on reliability, sufficient reliability was indeed attained. The multiple choice-essay correlations were obtained in two of the three studies GEDTS has conducted to date. In the 1983 study involving a national sample of high school seniors, the correlations between essay scores and scores on the Writing Skills test ranged from .55 to .69, depending on the essay topic (Swartz & Whitney, 1985). In the study involving GED examinees in Washington, D.C. and Iowa, the correlation between essay scores and Writing Skills test scores was .54. As indicated in Table 4, the higher correlations yielded in the high school senior study are primarily attributable to the greater variance of both essay and GED Writing Skills test scores among the high school senior sample. The extent of these correlations indicates that the two tests (multiple choice and essay) are measuring ***related but clearly different sets of skills***.

To give these coefficients some practical significance, this degree of correlation between the two measures is ***insufficient*** to permit the precise prediction of an essay score based on the multiple choice Writing Skills test score (or vice versa). That is, based on a regression analysis using these estimates of correlation, an examinee with a standard score of 40 on the GED Writing Skills test could be expected to earn a total essay score (on the 2 to 12 scale) anywhere from 2 to 6. As the scoring guide in Appendix A illustrates, there is a substantial difference between a paper scored as a 1 (by two different readers) and one scored as a 3 (by two different readers).

As the GED Language Arts Panel recommended, and as the research has demonstrated over the years (Godshalk, et al., 1966; White, 1985), the overall validity of a test of writing is improved significantly by combining highly reliable scores from an indirect measure of writing with a more valid (though relatively less reliable) direct measure of writing. In the case of the GED Writing Skills test, the combined scores will be reported as a single composite score on the same score scale as that used for all other tests in the GED Tests battery. GEDTS staff are currently investigating the possibility of varying the weighting of the essay score in the formation of the composite, depending on the reading reliability of the essay topic in use, so that the composite draws upon the reliability of the indirect measure as much as necessary to produce a composite writing score that has reliability comparable to that obtained for other tests in the battery (.81 to .94) (Whitney, et al., 1985).

Table 8
Comparison of Means and Standard Deviations
For High School Seniors and GED Examinees:
Essay Scores and Multiple Choice Standard Scores

	Essay Score			M.C. Standard Score		
Population	N	Mean	Standard Deviation	Mean	Standard Deviation	M.C./Essay Correlation
1983 High School Study[1]	747	6.5	2.4	50.8[2]	9.3	.69
All GED Examinees	1164	5.0	2.0			
D.C. Only	633	4.3	1.6	44.7	6.7	.54
Iowa Only	501	5.8	2.1	48.1	7.1	.48

Notes:

1. Persuasion topic only, since the persuasion topic was the closest in rhetorical form to those topics used with GED examinees.
2. Standard score means and standard deviations were based on the sample of 256 high school students who took both the Writing Skills multiple choice test and the persuasion topic.

Conclusions

The studies detailed in this report establish several important principles regarding the addition of an essay component to the GED Writing Skills test. First, an essay test will measure skills distinct from those currently measured in the multiple choice Writing Skills test, thus adding an additional degree of validity to the test. Second, essays written by GED examinees can be scored holistically with acceptable degrees of reliability. Moreover, sites using different modifications of the holistic scoring procedures, while maintaining the standard defined by a national norming group, can give approximately equal scores to essays regardless of the procedures in place; however, it is necessary to monitor ratings and individual scoring sessions to guard against shifts in the essay scoring scale.

These findings are particularly noteworthy in this study because in all cases readers were scoring essays representing somewhat restricted distributions of scores. As the mean scores for GED examinees indicate, the clear majority of essays were scored in the lower range of the score scale (1-3), regardless of where the papers were written or scored. This limited range of scores greatly affects the extent to which readers can be expected to score accurately, because throughout the scoring session they were seeing almost exclusively lower-half papers without the benefit of an occasional upper-half paper to serve the purpose of a comparison. Yet, despite this limitation, not to mention the "local reading" design in place for the Iowa readings, reader reliabilities were well within the range generally considered acceptable.

The chief implication of this finding is that it is technically feasible to include an essay component in the GED Writing Skills test and to rely upon a nationally-designed scoring guide and locally-implemented procedures to score the essays. While various practical issues remain to be addressed, the findings from these studies demonstrate that a nationally-defined and nationally-monitored essay scoring standard can be applied uniformly and consistently while allowing for scoring procedures to be adapted to meet the needs of local jurisdictions.

References

Applebee, A.N. (1981). *Writing in the secondary school: English and the content areas*. Urbana, IL: National Council of Teachers of English.

Bamberg, B. (1982). Multiple-choice and holistic essay scores: what are they measuring? *College Composition and Communication*, 404-406.

Boyer, E. L. (1983). *A report on secondary education in America*. Washington, D.C.: The Carnegie Foundation for the Advancement of Teaching.

Breland, H. M. (1983).*The direct assessment of writing skill: a measurement review*. New York: College Entrance Examination Board.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: a critical overview. *Research in the Teaching of English*, 18, 69-71.

Coffman W. E. (1971). Essay examinations. In R. L. Thorndike (Ed.), *Educational measurement*, (2nd ed.). Washington, D.C.: American Council on Education.

Conlan, G. (1976). *How the essay in the CEEB English test is scored*. Princeton, N.J.: Educational Testing Service.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. Cooper and L. Odell (Eds). *Evaluating writing: describing, measuring, judging*. Urbana, IL.: National Council of Teachers of English.

Education Commission of the States. (1984). *Direct Writing Assessments in the States*. (ECS Issuegram #50). Denver: ECS.

Educational EQuality Project. (1983). *Academic preparation for college: What students need to know and be able to do*. New York: The College Board.

Educational EQuality Project. (1984). *Academic preparation for the world of work*. New York: The College Board.

General Educational Development Testing Service. (1984, June). *GED tests specifications committee report*. Washington, D.C.: American Council on Education.

Godshalk, F. I., Swineford, F., and Coffman, W. E. (1966). *The measurement of writing ability*. New York: College Entrance Examination Board.

Goodlad, J.I. (1983). *A place called school*. New York: McGraw-Hill.

Iowa Department of Public Instruction. (1985, July). *Bright horizons: Iowa GED writing skills project final report*. Des Moines, IA: IDPI.

Lloyd-Jones, R. (1977). Primary trait scoring. In C. Cooper and L. Odell (Eds.), *Evaluating writing: describing, measuring, judging*. Urbana, Il.: National Council of Teachers of English.

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (1984). *High schools and the changing workplace: The employers' view*. (Report of the Panel on Secondary School Education for the Changing Workplace.) Washington, D.C.: National Academy Press.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for education reform*. Washington, D.C.: U.S Department of Education.

Spandel, V., & Stiggins, R. J. (1980). *Direct measures of writing skill: issues and applications*. Portland, OR.: Northwest Regional Educational Laboratory.

Spille, H. A. (1984). *Summary of surveys concerning future GED Tests (phase 2)*. (GED Testing Service Memo No. 31). Washington, D.C.: American Council on Education.

Swartz, R. (1985, July). *GED essay feasibility studies: report on practical issues*. (GED Testing Service, mimeograph). Washington, D.C.: American Council on Education.

Swartz, R., & Whitney, D. R. (1985, June). *The relationship between scores on the GED Writing Skills Test and on direct measures of writing*. (GED Testing Service, Research Studies, No. 6). Washington, D.C.: American Council on Education.

Task Force on Education for Economic Growth. (1983). *Action for excellence*. Denver: Education Commission of the States.

White, E. M. (1984). Holisticism. *College Composition and Communication*, 35, 405-409.

White, E. M. (1985). *Teaching and assessing writing*. San Francisco: Jossey-Bass.

Whitney, D.R., Malizio, A.G., and Patience, W.M. (1985, May). *The reliability and validity of the GED Tests*. (GED Testing Service, Research Brief No. 6). Washington, D.C: American Council on Education.

Appendix A

GED ESSAY SCORING GUIDE
April, 1985

Papers will show some or all of the following characteristics.

Upper-half papers make clear a definite purpose, pursued with varying degrees of effectiveness. They also have a structure that shows evidence of some deliberate planning. The writer's control of English usage ranges from fairly reliable at 4 to confident and accomplished at 6.

6 Papers scored as a 6 tend to offer sophisticated ideas within an organizational framework that is clear and appropriate for the topic. The supporting statements are particularly effective because of their substance, specificity, or illustrative quality. The writing is vivid and precise, though it may contain an occasional flaw.

5 Papers scored as a 5 are clearly organized with effective support for each of the writer's major points. The writing offers substantive ideas, though the paper may lack the flair or grace of a 6 paper. The surface features are consistently under control, despite an occasional lapse in usage.

4 Papers scored as a 4 show evidence of the writer's organizational plan. Support, though sufficient, tends to be less extensive or convincing than that found in papers scored as a 5 or 6. The writer generally observes the conventions of accepted English usage. Some errors are usually present, but they are not severe enough to interfere significantly with the writer's main purpose.

Lower-half papers either fail to convey a purpose sufficiently or lack one entirely. Consequently, their structure ranges from rudimentary at 3, to random at 2, to absent at 1. Control of the conventions of English usage tends to follow this same gradient.

3 Papers scored as a 3 usually show some evidence of planning or development. However, the organization is often limited to a simple listing or haphazard recitation of ideas about the topic, leaving an impression of insufficiency. The 3 papers often demonstrate repeated weaknesses in accepted English usage and are generally ineffective in accomplishing the writer's purpose.

2 Papers scored as a 2 are characterized by a marked lack of development or inadequate support for ideas. The level of thought apparent in the writing is frequently unsophisticated or superficial, often marked by a listing of unsupported generalizations. Instead of suggesting a clear purpose, these papers often present conflicting purposes. Errors in accepted English usage may seriously interfere with the overall effectiveness of these papers.

1 Papers scored as a 1 leave the impression that the writer has not only not accomplished a purpose, but has not made any purpose apparent. The dominant feature of these papers is the lack of control. The writer stumbles both in conveying a clear plan for the paper and in expressing ideas according to the conventions of accepted English usage.

0 The zero score is reserved for papers which are blank, illegible, or written on a topic other than the one assigned.

Appendix B

HOW THE GED ESSAY IS SCORED

GED Testing Service

Holistic Scoring

Each paper written for the GED Writing Skills Test is scored by two trained readers. The readers score the papers "holistically;" that is, the piece of writing is evaluated on the basis of its *overall* effectiveness. The salient features of holistic scoring are well described by Conlan (1976):

> The basic assumptions of the holistic reading are that each of the factors involved in writing skill is related to all the others and that no one factor can be separated from the others. Readers must judge each essay as a whole; they must read each paper for the impression its totality makes. A misspelled word, a comma splice, a sentence fragment, a misplaced modifier should carry no great weight in scoring a paper. The candidate is entitled to make some mistakes: he or she is writing hurriedly in a tense situation, without recourse to a dictionary, without the customary time for deliberation. If readers read each paper as a whole, then, they are better able to judge the competence of the writer. If the paper is poorly written, that will be part of the reader's first impression; there is no need to analyze it word by word to decide that it is badly done. If the paper contains mistakes because the writer, though capable, has been forced to hurry, the reader will judge the general quality of the paper most effectively by reading it as a whole. The motto of the reading might well be "Read quickly and judge. Do not reread."

The Six Point Scale

The scale used by each reader in the holistic evaluations ranges from one (low) to six (high). The two readers' scores are then added, resulting in a range of scores from two to twelve. If the two readers' scores differ by more than one point, the paper is scored a third time. The total score for papers scored by three readers is twice the average of the three scores.

Because the six point scale is an even-numbered scale, there is no midpoint. The lack of a midpoint forces readers away from a natural tendency to drift towards the middle. With an even-numbered scale, readers must decide whether each paper belongs in the upper half or the lower half of the papers being scored. Readers can make this decision only if they have some awareness of the total range of writing ability represented by the papers; this is provided by sample papers in conjunction with a descriptive scoring guide.

How the Standards are Defined

The standards for an essay scoring session are defined by the GED essay scoring guide and by sample papers illustrating the different points on the scoring scale. These standards were originally established by examining papers written by a national sample of high school seniors and will be maintained by following the same process for each topic that is used in the GED Writing Skills Test. A group of highly experienced readers reads papers in the pool of high school essays to establish the range of abilities demonstrated by the norming population of (high school) students. After they have selected papers which represent the full range of ability, the readers assign the top score to the best of the papers, the lowest score to the weakest papers, and corresponding scores to other papers throughout the range. The readers' judgments during this process are made independently, and the papers that ultimately serve as samples include only those on which the the readers' scores agree. These sample papers represent the full range of abilities demonstrated by a national sample of high school seniors and thus serve as standard setters, or "rangefinders."

The GED Essay Scoring Guide

The GED essay scoring guide is a further articulation of these standards and accompanies the rangefinders in training sessions. Because it was developed inductively, the scoring guide is descriptive rather than prescriptive. That is, after the readers selected the rangefinders, they attempted to describe, in general terms, the characteristics of papers at different points on the scoring scale. While the sample papers illustrate standards only for the specific topic on which they are written, the scoring guide defines the characteristics papers should exhibit regardless of the topic the paper is written on. Although there is ample research demonstrating that different topics exhibit different strengths and weaknesses of a writer, a generic guide can be used if topics are developed with the scoring guide in mind. All topics developed for use in the GED Essay test are of the same rhetorical type and are of approximately equal length, reading level, and format. In addition, all potential topics are field tested and scored by trained readers, using the GED scoring guide to define standards. If readers find that papers written on a potential topic cannot be scored using the scoring guide, or that papers consistently

exhibit traits other than those described in the guide, the topic is rejected. A further check on the match of topic to scoring guide is made through a statistical analysis wherein the mean and variance of examinee scores on potential topics must approximate the mean and variance of the anchor topic for which the scoring guide was written.

Training of Readers

While the specific procedures for training readers and conducting the essay scoring session depend upon the number of papers to be scored and the number of readers participating, many of the basic principles remain the same regardless of the size of the effort. Readers are usually trained prior to or at the beginning of the essay scoring session. After an introduction to the principles of holistic scoring provided by a chief reader, readers are given the topic, the scoring guide, and the rangefinders, which include at least one paper at each point on the score scale. After a discussion of the tasks required by the topic, and a review of the qualities enumerated on the scoring guide, readers are asked to read the entire set of rangefinders, evaluating them swiftly on the basis of an overall impression and ranking them from best to poorest. Because the set of rangefinders includes more papers than there are points on the score scale (i.e., 8 papers for a six point scale), readers are instructed that they must use the entire range of scores. This is an important instruction in the training because many readers are reluctant to award scores at the top or at the bottom of the score scale. If readers are directed to give the top score to the best paper in the set of rangefinders, they are more likely to acknowledge that a "6" paper does not—and need not—represent perfection.

In a large group training session, the chief reader calls for a show of hands to indicate how readers scored the rangefinders. Ideally, all of the readers will have scored each of the rangefinders identically. In some readings, involving highly experienced readers, this ideal is realized. However, with inexperienced readers there are often substantial differences among readers' scores at this point. To help resolve those differences, readers designated as "table leaders" conduct discussions among the four to six readers at their tables to attempt to bring each individual to a point of consensus with the group. At the heart of the holistic scoring process is the necessity for each reader to be willing to accept—and in some cases consciously shift to—the standards defined in the scoring guide and adopted by the group. Following the discussion of the rangefinders, additional work is done towards this goal through the scoring of more sample papers, distributed singly or in sets of two or three. The scoring and discussion of sample papers continues until the entire group of readers begins to show a consensus in their scoring. At this point, the training period ends and the actual scoring of essays begins, although table leaders may continue to work with individual readers if necessary.

Reliability and Stability

The goals of an essay scoring session are inter-rater reliability and reading "stability." Inter-rater reliability is the degree to which readers agree with each other; generally, the fewer papers which require a third reading, the greater the inter-rater reliability, though this index is affected by other variables as well. Reading stability is the degree to which papers are scored according to the fixed standards described in the scoring guide. This is a particularly important feature of an essay scoring session because the standards for scoring GED essays must remain fixed, regardless of when the essay is administered, where it is scored, or what specific procedures were used in the scoring session itself. It is important to note that a high degree of inter-rater reliability does not ensure reading stability. In short, just because readers are agreeing with each other on essay scores does not mean that they are assigning the scores according to the standards defined on the scoring guide.

To achieve these two goals, the reinforcement of scoring standards continues well after the initial training session is over. As the readers score papers, a table leader (in a small reading) selects scored papers at random to verify that the scoring is consistent with the definitions in the scoring guide. In cases of a disagreement, the chief reader or table leader discusses the paper with the reader. The monitoring process continues throughout the entire scoring session, and, in large readings, the chief reader periodically reviews the scoring of the table leaders.

When scoring large numbers of papers, readers generally read for periods from 30 to 45 minutes at a time and then take breaks. At the beginning of each reading period, additional sample papers are distributed and scored as an additional check on reader accuracy. Generally, readers score between 30 and 50 papers per hour, depending on various factors such as the reader's experience, the length of the papers, and the time of day. Readers are encouraged to read swiftly because a slow, deliberate reading of a paper tends towards analytical, rather than holistic, evaluation of writing. But readers are frequently reminded that accuracy is far more important than speed. Through this system of checks and re-checks, some assurance is gained that all readers are scoring according to the standards defined by the scoring guide and not according to one defined by each individual.

ACE Board of Directors

Executive Committee

Philip H. Jordan, Jr., President
Kenyon College, *Chair*

Frank H. T. Rhodes, President
Cornell University, *Vice Chair*

Ronald W. Roskens, President
University of Nebraska, *Immediate Past Chair*

Joyce S. Tsunoda
Chancellor for Community Colleges
University of Hawaii, *Secretary*

Robert H. Atwell, President
American Council on Education

Stelle Feuers, President
Los Angeles City College

Janet D. Greenwood, President
Longwood College

Association Representatives

Margaret W. Bridwell, M.D.
Director of the Health Center
University of Maryland

Robert V. Cramer, President
Carroll College

The Reverend Charles Currie, S.J., President
Xavier University

Judith S. Eaton, President
Commmunity College of Philadelphia

Brother Patrick Ellis, F.S.C., President
La Salle University

Carol J. Guardo, Provost
University of Hartford

Joab M. Lesesne, Jr., President
Wofford College

Aubrey K. Lucas, President
University of Southern Mississippi

Ellis E. McCune, President
California State University, Hayward

Luna I. Mishoe, President
Delaware State College

John W. Ryan, President
Indiana University System

John S. Toll, President
University of Maryland

Theodore Ziolkowski, Dean
The Graduate School
Princeton University

Class of 1986

Alfredo G. de los Santos Jr.
Vice Chancellor for Educational Development
Maricopa Community Colleges

Harold W. Eickhoff, President
Trenton State College

Stelle Feuers, President
Los Angeles City College

Stanley O. Ikenberry, President
University of Illinois

Eamon M. Kelly, President
Tulane University

Betty Lentz Siegel, President
Kennesaw College

Class of 1987

Roscoe C. Brown, Jr., President
Bronx Community College
City University of New York

William M. Fulkerson, Jr., President
Adams State College

Janet D. Greenwood, President
Longwood College

William J. Maxwell, President
Jersey City State College

Patsy H. Sampson, President
Stephens College

James J. Whalen, President
Ithaca College

Class of 1988

Robert L. Albright, President
Johnson C. Smith University

David W. Breneman, President
Kalamazoo College

Edward T. Foote II, President
University of Miami

William P. Gerberding, President
University of Washington

Bruce E. Whitaker, President
Chowan College

Brunetta R. Wolfman, President
Roxbury Community College

11/1/85